인센티브 메커니즘 기반 Non-IID 사용자를 위한 프라이버시 보호 연합학습

김범준, 서효운*, 최완
서울대학교 전기정보공학부, 뉴미디어통신공동연구소, *광운대학교 전자통신공학과

eithank96@snu.ac.kr, *hyowoonseo@kw.ac.kr, wanchoi@snu.ac.kr

Private Federated Learning for Non-IID Clients Based on Incentive Mechanism

Bumjun Kim, Hyowoon Seo*, Wan Choi
Department of Electrical and Computer Engineering, INMC, Seoul National University,

*Department of Electronics and Communications Engineering, Kwangwoon University.

요 약

연합학습 (federated learning)은 분산된 여러 사용자의 데이터를 활용하고 사용자가 중앙서버와 데이터를 직접 주고받지 않아 프라이버시를 보장하는 장점을 가지고 있다. 그러나 중앙서버에서 모델 전도 공격 (model inversion attack) 및 회원 추론 공격 (membership inference attack)과 같은 공격이 가능하다는 사실이 알려지면서 연합학습에서 추가적으로 프라이버시를 보장하기위한 연구가 활발히 이루어지고 있다. 차분 프라이버시 (differential privacy) 기술은 프라이버시를 정량적으로 측정하여 프라이버시 보호 정도를 측정하고 가우시안 메커니즘 (Gaussian mechanism)과 같은 방식으로 프라이버시를 보호하는 기술이다. 또한, 연합학습은 사용자의 데이터를 활용하여 학습하므로 사용자의 데이터 분포에 따라 학습 성능이 크게 좌우된다. 따라서 본 연구에서는 중앙서버가 연합학습에 참여하는 사용자들의 데이터를 궁금해하는 환경에서, non-independent and identically distributed (non-IID)한 데이터를 가지고 있는 여러 사용자가 인센티브 메커니즘 (incentive mechanism)을 통해 학습 성능을 높이는 동시에 최대한의 프라이버시를 보장할 수 있는 기법을 제안한다.

Ⅰ. 서 론

최근, 사용자의 프라이버시를 보장하고, 많은 양의 데이터를 사용하여 학습하는 연합학습이 딥러닝 분야에 널리 사용되고 있다. 연합학습은 모델 학습을 위해 중앙서버가 데이터를 직접 활용하는 형태가 아니며, 분산된 사용자가 중앙서버와 데이터가 아닌 국소 그라디언트 (local gradient)를 교환함으로써 프라이버시를 보장하는 장점이 있다. 하지만 실제 환경에서 연합학습을 활용했을 때 사용자가 가지고 있는 데이터가 non-IID 할 경우 학습 성능이 크게 저하된다는 문제점이 있고, 그라디언트 및 모델을 통하여 데이터를 추출하는 모델 전도 공격, 특정 데이터가 학습에 사용된 데이터인지 아닌지를 알아내는 회원 추론 공격으로 인해 연합학습에서도 프라이버시가 침해될 수 있다는 사실이 알려졌다.

이와 관련해 본 연구에서는 non-IID 한 데이터를 가지고 있는 여러 사용자가 연합학습을 하는 상황에서 중앙서버로부터 데이터를 제공받는 대신 중앙서버에 대한 프라이버시 보장 정도를 인센티브 메커니즘을 통하여 결정하는 기법을 제안한다. 제안하는 기법을 통해 사용자는 데이터의 non-IID 문제를 해결하는 동시에 프라이버시 유출을 최소화한다.

Ⅱ. 본론

가. 시스템 모델

본 논문에서는 over-the-air 환경에서 중앙서버가 가지고 있는 데이터의 크기가 D_g 이고, 각 D_k 만큼의 non-IID 한 데이터를 가지고 있는 총 K 명의 사용자가 참여하는 환경을 고려한다. 사용자는 프라이버시를 보장하기 위해 그라디언트에 노이즈를 주입하며, 노이즈의 양은 논문 [1]에서 구한 프라이버시에 따른 노이즈인 $\sum_k |h_k|^2 \beta_k P_k + \sigma_m^2$ 로 해석할 수 있다. 또한, 데이터의 non-IID 한 문제를 해결하기 위해 서버는 사용자에게 공통 데이터를 나눠주고 [2], 대신 프라이버시에 대한 침해 권리를 얻게 된다. 각 사용자가 매 상향 링크 (uplink) 때 중앙서버에 자신의 그라디언트를 전송할 확률을 p_k 로 정의하며 사용자와 중앙서버는 채널을 알고 있다고 가정한다.

나. 문제 정립

프라이버시와 학습 성능을 동시에 최적화할 수 있는 최적의 공통 데이터의 양을 결정하기 위해 문제 **P** 는 다음과 같이 정립할 수 있다.

(P)

$$\max \mu = f(x) - p_m \rho$$

$$\text{s.t.} \quad x \le D_g$$

$$\rho \ge c$$

이때, $f(x)$ 는 중앙서버가 사용자에게 전송하는 공통 데이터 x에 따른 학습 성능 개선 정도를 나타내는 함수, p_m 은 사용자가 서버에게 전송하는 최소 확률 값, ρ 는 프라이버시를 보장하는 정도, c 는 사용자가 최소한으로 보장해야 하는 프라이버시 정도를 나타낸다. $f(x)$ 와 ρ 는 중앙서버가 $(f(x), \rho)$ 의 형태로 가지고 있으며 이는

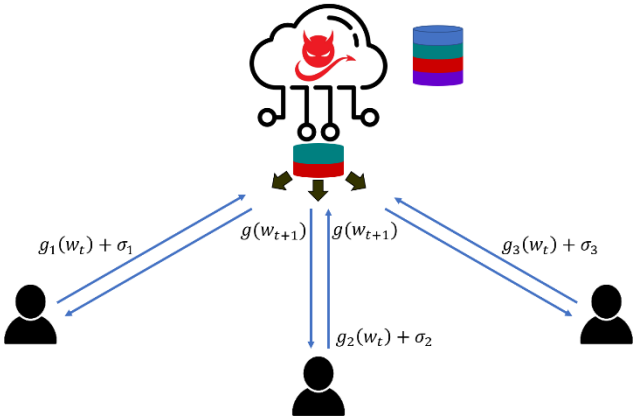

그림 1. Non-IID 데이터를 소유한 사용자를 위한 프라이버시 보호 연합학습 기법 예시.

프라이버시 보장 정도에 따라 중앙서버가 제공할 수 있는 공통 데이터에 따른 성능 개선 정도를 의미한다.

문제 **P**는 (f(x), ρ)가 주어졌을 때 얻을 수 있는 이득 μ 를 최대화하는 것을 목표로 하며 본 연구에서는 문제 **P**의 해를 exhaustive search 를 통해 구했다. 중앙서버는 문제 **P** 를 통해 구한 x, p_m, ρ 을 사용자에게 전달하고 사용자는 공통 데이터 x 를 전달받는 대신 ρ 만큼의 프라이버시를 보장하도록 그라디언트 신호를 전송한다. 이때, 각 사용자가 매 상향링크 때 중앙서버에 그라디언트를 전송할 확률은 $p_k = p_m(D_k/(D_k + x))$ 이고 이는 공통 데이터 대비 개인 데이터가 많은 사용자가 그렇지 않은 사용자 대비 프라이버시에 취약하다는 점을 보완하기 위해 가중치를 준 것으로 해석할 수 있다.

다. 시뮬레이션

본 시뮬레이션에서는 MNIST 데이터셋을 활용했으며, 분할된 사용자는 두 개의 convolution layer 와 3 개의 fully connected layer 로 구성하였고, 총 200 번의 라운드를 실험하였다. 각 데이터셋에 대하여 1) 중앙서버의 도움을 받지 않은 상황에서 프라이버시를 보장하는 경우, 2) 제안된 기법을 사용한 경우, 3) 중앙서버의 도움을 받지 않은 상황에서 프라이버시 보장을 하지 않은 경우로 총 세 가지의 실험하였다.

표 1 을 통해, 제안된 기법은 중앙서버의 도움으로 학습 성능이 개선되었음을 확인할 수 있었다. 이는 중앙서버로부터 전달받는 데이터로 인해 사용자 데이터의 non-IID 특징이 줄어들고, 사용자가 중앙서버와의 약속을 위해 프라이버시를 위한 노이즈의 양을 줄여 신호를 전송하기 때문이다.

분석 / 데이터셋	정확도		
	실험 1	실험 2	실험 3
MNIST	90.95	96.76	92.68
CIFAR-10	55.66	68.72	63.68

표 1. 데이터셋 및 실험 별 정확도 분석.

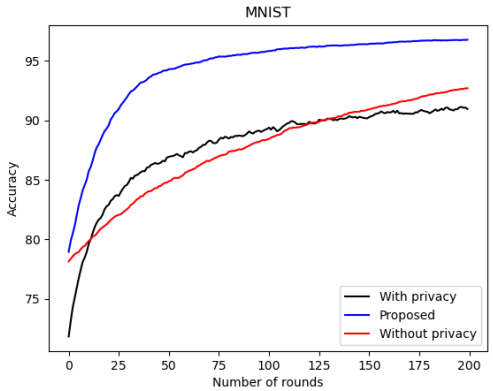

그림 2. MNIST 데이터셋을 활용한 실험 별 정확도.

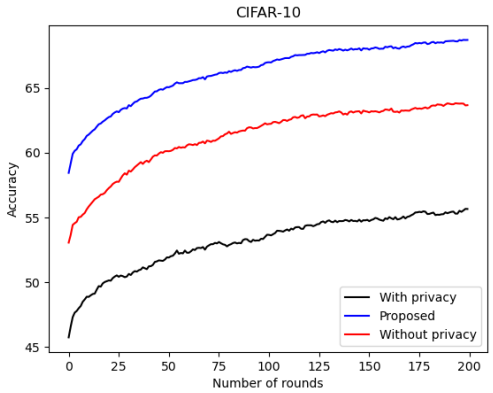

그림 3. CIFAR-10 데이터셋을 활용한 실험 별 정확도.

Ⅲ. 결론

본 논문에서는 non-IID 한 데이터를 가지고 있는 여러 사용자가 연합학습을 하는 상황에서 사용자는 데이터의 non-IID 문제를 해결하는 동시에 프라이버시 유출을 최소화하는 기법을 제안하였다. 중앙서버는 가지고 있는 공통 데이터를 분할되어 있는 사용자에게 전송하여 사용자의 데이터셋의 non-IID 정도를 해결하는 대신, 사용자는 프라이버시를 지키기 위해 전송했던 노이즈의 양을 줄이는 기법을 제안하였다. 공통 데이터의 양과 노이즈 양에 대한 관계식을 이득에 대한 식으로 정리한 후, 해당 이득을 최대화하여 사용자가 좋은 성능 결과 및 프라이버시 유출을 최소화하는 기법을 제안하였다.

ACKNOWLEDGMENT

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2021R1A2C2003230).

참 고 문 헌

[1] M. Seif, R. Tandon and M. Li, "Wireless Federated Learning with Local Differential Privacy," 2020 IEEE International Symposium on Information Theory (ISIT), pp. 2604-2609, 2020.

[2] Y. Zhao, M. Li, L. Lai, N. Suda, D.Civin, and V. Chandra, "Federated learning with non-iid data," arXiv preprint arXiv:1806.00582, 2018.